Estimation of Fixed Effects in a 2X2 Factorial Experiment with Unequal Subclass Numbers*

BU-196-M Henry E. Schaffer May, 1965

Least Squares estimators and the variances of these estimators are given for the effects in a 2X2 factorial experiment with unequal subclass numbers. Two models are considered, one with and one without interaction.

*This work was supported by a postdoctoral Traineeship in the Genetics Training Program under Grand TI GM 1035 from the National Institute of General Medical Sciences.

# Estimation of Fixed Effects in a 2X2 Factorial Experiment with Unequal Subclass Numbers*

BU-196-M Henry E. Schaffer May, 1965

## Model Including Interaction

When the effects of two treatments applied singly and jointly are studied there are several ways of expressing the general model. Table I gives a full model for the expected values of observations.

Table I

| | | Treatment A Absent | Treatment A Present |
|---|---|---|---|
| Treatment B | Absent | m | m+a |
| | Present | m+b | m+a+b+c |

m is the "control" value or the expected value in the absence (or at the lower levels) of both treatments.

a is the effect of Treatment A

b is the effect of Treatment B

c is the effect of the interaction of Treatments A and B

It is desired to estimate m, a, b, c and to compute the standard errors of estimators for use in tests of signifficance and construction of confidence intervals.

Table II gives the number of observations and the mean for each cell. There must be at least one observation in each cell to estimate all the effects.

* This work was supported by a postdoctoral Traineeship in the Genetics Training Program under Grand TI GM 1035 from the National Institute of General Medical Sciences.

Table II

| | | Treatment A | |
|---|---|---|---|
| | | Absent | Present |
| Treatment | Absent | i obsvns., mean $\bar{Y}_{00}$ | j obsvns., mean $\bar{Y}_{A0}$ |
| B | Present | k obsvns., mean $\bar{Y}_{0B}$ | n obsvns., mean $\bar{Y}_{AB}$ |

The Least Squares estimators of the effects in this model and the variances of these estimators are:

$$\hat{m} = \bar{Y}_{00} \qquad \sigma^2_{\hat{m}} = \left(\frac{1}{i}\right)\sigma^2$$

$$\hat{a} = \bar{Y}_{A0} - \bar{Y}_{00} \qquad \sigma^2_{\hat{a}} = \left(\frac{1}{i} + \frac{1}{j}\right)\sigma^2$$

$$\hat{b} = \bar{Y}_{0B} - \bar{Y}_{00} \qquad \sigma^2_{\hat{b}} = \left(\frac{1}{i} + \frac{1}{k}\right)\sigma^2$$

$$\hat{c} = \bar{Y}_{AB} - \bar{Y}_{A0} - \bar{Y}_{0B} + \bar{Y}_{00} \qquad \sigma^2_{\hat{c}} = \left(\frac{1}{i} + \frac{1}{j} + \frac{1}{k} + \frac{1}{n}\right)\sigma^2$$

where $\sigma^2$ is the within cell variance.

The usual estimate of $\sigma^2$, called $s^2$, is the pooled within cell sample variance. This is the sum of the corrected sums of squares of observations within each cell divided by the sum of the degrees of freedom in each cell. Symbolically

$$s^2 = \frac{SS(00) + SS(A0) + SS(0B) + SS(AB)}{\text{d.f. in } (00) + \text{d.f. in } (A0) + \text{d.f. in } (0B) + \text{d.f. in } (AB)} \quad - - - (1)$$

and has $(i - 1) + (j - 1) + (k - 1) + (n - 1)$ d.f. when there are two or more observations in each cell. If a cell has less than two observations it is dropped from the calculation of $s^2$ in formula (1).

When $s^2$ is substituted for $\sigma^2$ in the formula for the variance of an estimator, and the square root is taken to obtain this standard error, the number of degrees of freedom in this standard error is the number of degrees of freedom in $s^2$. This is also the degrees of freedom used in entering the t-table for tests of significance and construction of confidence intervals. For example, if there are m degrees of freedom in $s^2$, the t test of the null hypotheses a = 0 is

$$t = \frac{\hat{a}}{s_a} = \frac{\hat{a}}{\sqrt{(\frac{1}{i} + \frac{1}{j})s^2}} .$$

This t would then be considered significant (at the 5% level) if it were larger than the tabular value $t_{.05}$, m d.f.

<u>Case of $i = j = k = n = 1$.</u>

In this case $s^2$ can not be computed within cells. While all the effects can be estimated, they can't be tested without an estimate of $\sigma^2$. It is only possible to estimate $\sigma^2$ from the data if it can be assumed that one of the effects estimated has a true value of zero. Then the value of that estimate only reflects random variation.

If, as is most common, it is the interaction that is assumed to be non-existent, then $\hat{c}$ only reflects random variation and $\hat{c}^2$ is an estimate of $(\frac{1}{i} + \frac{1}{j} + \frac{1}{k} + \frac{1}{n})\sigma^2 = 4\sigma^2$. Therefore the $s^2$ we can use is

$$s^2 = \frac{1}{4}\hat{c}^2$$

with one degree of freedom.

If there actually is interaction, it will bias upwards this estimate of $\sigma^2$ and so will tend to make the confidence intervals too long and the tests less likely to show significance.

This use of $\hat{c}^2$, as an estimate of the variance, in testing the other effects is the same as using the interaction mean square in the analysis of variance to test the main effects. Actually all the analyses in this paper could be put in the form of analysis of variance, but when the estimates of the effects are of interest it is more appropriate to test them directly.

<u>Model Without Interaction.</u>

The estimates of m, a, b obtained using the full model (including interaction) are unbiased and the tests are still correct when there is no interaction (i.e. when c = 0).

However when it is known that there is no interaction, this knowledge can be used to give better estimators (they have smaller variances) and more powerful tests. If there is some doubt about this assumption, then the methods of the previous sections are preferable. The presence of interaction will cause the estimators given in this section to be biased.

When there is no interaction, estimation of all the effects is still possible even if one of the cells has no observations, but the estimators given in this section are for the case of observations in all cells. As before, only cells with two or more observations can be used in formula (1) to calculate $s^2$.

The model without interaction is shown in Table III. The experiment is as described in Table II.

Table III

| | | Treatment A | |
|---|---|---|---|
| | | Absent | Present |
| Treatment | Absent | m | m + a |
| B | Present | m + b | m + a + b |

The Least Squares estimators of the effects and the variances of these estimators are:

$$\hat{m} = \frac{1}{t}\left[\left(\frac{1}{j} + \frac{1}{k} + \frac{1}{n}\right)\bar{Y}_{OO} + \frac{1}{i}\left(\bar{Y}_{AO} + \bar{Y}_{OB} - \bar{Y}_{AB}\right)\right] \qquad \sigma^2_{\hat{m}} = \frac{1}{t}\left(\frac{1}{i}\right)\left(\frac{1}{j} + \frac{1}{k} + \frac{1}{n}\right)\sigma^2$$

$$\hat{a} = \frac{1}{t}\left[\left(\frac{1}{k} + \frac{1}{n}\right)\left(\bar{Y}_{AO} - \bar{Y}_{OO}\right) + \left(\frac{1}{i} + \frac{1}{j}\right)\left(\bar{Y}_{AB} - \bar{Y}_{OB}\right)\right] \qquad \sigma^2_{\hat{a}} = \frac{1}{t}\left(\frac{1}{k} + \frac{1}{n}\right)\left(\frac{1}{i} + \frac{1}{j}\right)\sigma^2$$

$$\hat{b} = \frac{1}{t}\left[\left(\frac{1}{j} + \frac{1}{n}\right)\left(\bar{Y}_{OB} - \bar{Y}_{OO}\right) + \left(\frac{1}{i} + \frac{1}{k}\right)\left(\bar{Y}_{AB} - \bar{Y}_{AO}\right)\right] \qquad \sigma^2_{\hat{b}} = \frac{1}{t}\left(\frac{1}{j} + \frac{1}{n}\right)\left(\frac{1}{i} + \frac{1}{k}\right)\sigma^2$$

where $t = \left(\frac{1}{i} + \frac{1}{j} + \frac{1}{k} + \frac{1}{n}\right)$.

The pooled within cell variance, $s^2$, can be computed as before from formula (1) and used as an estimate of $\sigma^2$.

There is another estimate of $\sigma^2$ which may be obtained since the four cell means are only used to estimate three parameters. This is the sum of squares of deviations (from the model):

$$i(\bar{Y}_{00} - \hat{m})^2 + j(\bar{Y}_{A0} - \hat{m} - \hat{a})^2 + k(\bar{Y}_{0B} - \hat{m} - \hat{b})^2 + n(\bar{Y}_{AB} - \hat{m} - \hat{a} - \hat{b})^2$$

with one degree of freedom. This can be used alone as $s^2$ when it is not possible to use formula (1), or can be pooled with the within cell variance by adding it to the numerator in formula (1) and adding one degree of freedom to the denominator. In practice, this extra estimate is ignored when there are more than about 15 d.f. in $s^2$.

Tests of significance and confidence limits can then be calculated using the t-table.

## Covariances of the Estimators

The covariances of the estimators are needed, in addition to the variances which are given above, in computing the variance of linear functions of the estimators. (e.g., if a - b is of interest, $\mathrm{Var}(\hat{a} - \hat{b}) = \mathrm{Var}\ \hat{a} + \mathrm{Var}\ \hat{b} - 2\mathrm{Cov}(\hat{a},\hat{b})$.)

For the model including interaction the covariances are:

$$\begin{array}{cc} & \begin{array}{ccc} \hat{a} & \hat{b} & \hat{c} \end{array} \\ \begin{array}{c} \hat{m} \\ \hat{a} \\ \hat{b} \end{array} & \left[\begin{array}{ccc} -\frac{1}{i} & -\frac{1}{i} & \frac{1}{i} \\ & \frac{1}{i} & -\frac{1}{i} - \frac{1}{j} \\ & & -\frac{1}{i} - \frac{1}{k} \end{array}\right] \sigma^2 . \end{array}$$

For the model without interaction the covariances are:

$$\begin{array}{cc} & \begin{array}{cc} \hat{a} & \hat{b} \end{array} \\ \begin{array}{c} \hat{m} \\ \hat{d} \end{array} & \left[\begin{array}{cc} -\frac{1}{i}(\frac{1}{k} + \frac{1}{n}) & -\frac{1}{i}(\frac{1}{j} + \frac{1}{n}) \\ & (\frac{1}{in} - \frac{1}{jk}) \end{array}\right] \frac{\sigma^2}{t} . \end{array}$$